

Tabla de Contenido

1. Introducción	1
1.1. Antecedentes	2
1.1.1. Web Intelligence Centre - WIC	2
1.1.2. Proyecto OpinionZoom	2
1.2. Descripción y Justificación	5
1.3. Objetivos	6
1.3.1. Objetivo General	6
1.3.2. Objetivos Específicos	7
1.4. Hipótesis de investigación	7
1.5. Alcances	7
1.6. Resultados Esperados	8
1.7. Metodología	8
1.8. Estructura del Informe	10
2. Marco Teórico	11
2.1. Web 2.0	11
2.2. Twitter	12
2.2.1. APIs de Twitter	13
2.3. Knowledge Database Discovery - KDD	14
2.3.1. Data Mining	16
2.4. Text Mining	17
2.4.1. Procesamiento de texto	17
2.4.2. Bag-Of-Words	20
2.5. Machine Learning	21
2.5.1. Algoritmos Supervisados	23
2.5.2. Multinomial Naive Bayes	24
2.5.3. Multi-Class Support Vector Machines (SVM)	24
2.5.4. Artificial Neural Networks	26
2.5.5. Multinomial Logistic Regression	29
2.5.6. Decision Trees	30
2.6. Evaluación de Resultados para modelos de clasificación	31
2.6.1. Métricas de Desempeño	31
2.6.2. Métricas de Acuerdo	33
2.6.3. K-Fold Cross Validation	34
2.7. Modelo de Tópicos	34
2.7.1. Latent Dirichlet Allocation	35

2.8.	Reclamo	37
2.8.1.	Definición	37
2.8.2.	Reclamos en Twitter	38
2.8.3.	Reclamos en la Empresa	38
3.	Modelos de Clasificación de tópicos en Twitter	40
3.1.	Investigación de Bharath, Sriram, et al.	40
3.1.1.	Set de Datos	40
3.1.2.	Tópicos	41
3.1.3.	Procesamiento de Texto	41
3.1.4.	Modelos y Desempeño	42
3.1.5.	Conclusiones	42
3.2.	Investigación de Fernández Anta, Antonio, et al.	43
3.2.1.	Set de Datos	43
3.2.2.	Tópicos	43
3.2.3.	Procesado del Texto	43
3.2.4.	Modelos y Desempeño	44
3.2.5.	Conclusiones	45
3.3.	Investigación de Batista, Fernando, et al.	45
3.3.1.	Set de Datos	45
3.3.2.	Tópicos	45
3.3.3.	Procesado de Texto	45
3.3.4.	Modelos y Desempeño	46
3.3.5.	Conclusiones	46
3.4.	Investigación de Ebert, Sebastian, et al.	47
3.4.1.	Set de Datos	47
3.4.2.	Tópicos	48
3.4.3.	Procesado de Texto	48
3.4.4.	Modelos y Desempeño	49
3.4.5.	Conclusiones	49
3.5.	Resumen y Elección	50
4.	Diseño y Construcción del set de datos	51
4.1.	Elección de Rubro	51
4.1.1.	Industria Financiera	53
4.1.2.	Industria de Locales Comerciales	55
4.1.3.	Industria de Telecomunicaciones	56
4.1.4.	Resumen y Elección del Rubro a Utilizar	58
4.2.	Elección de Empresas	59
4.2.1.	Resumen y Elección de Empresas a Utilizar	62
4.3.	Dinámica en Twitter	63
4.4.	Selección y Extracción de Datos	64
4.4.1.	REST API - Search Keyword	66
4.4.2.	Base de Datos <i>La Gorda</i>	66
4.4.3.	Crawler de Usuarios Chilenos	68
4.4.4.	Resumen del Proceso de Extracción de Datos	69
4.5.	Tamaño y Selección del Set de Entrenamiento	69

4.6.	Definición de Categorías	70
4.7.	Etiquetado del set de datos	72
4.7.1.	Diseño del proceso de etiquetado	72
4.7.2.	Resultados del proceso de etiquetado	75
4.7.3.	Resumen	76
5.	Modelo de Detección y Clasificación de Reclamos	78
5.1.	Especificaciones de la Implementación	79
5.2.	Detección de Reclamos	79
5.2.1.	Modelo Bag-Of-Words	80
5.2.2.	Resumen y Elección	89
5.2.3.	Datos Ambiguos	90
5.3.	Categorización de Reclamos	91
5.3.1.	Modelo Bag-Of-Words	91
5.3.2.	Resumen y Elección	98
5.3.3.	Datos Ambiguos	99
6.	Diseño e Integración del Módulo de Reclamos	101
6.1.	Clasificador de reclamos	101
6.1.1.	API de reclamos (RAPI)	101
6.2.	Integración del clasificador de reclamos	102
6.2.1.	Algoritmo de clasificación de tweets	104
6.2.2.	Algoritmo de agrupación de tweets	105
6.3.	Diseño del Módulo de Reclamos en el sitio OpinionZoom	105
6.3.1.	Módulo de Reclamos - Sección Reclamos	107
6.3.2.	Módulo de Reclamos - Sección Categorías	110
6.4.	Nube de Palabras	110
6.5.	Validación con clientes	114
7.	Conclusiones	116
7.1.	Conclusiones	116
7.2.	Trabajo futuro	118
	Bibliografía	121
	Anexos	128
	A. Lista de Stopwords	129
	B. Gráficos Módulo de Reclamos - Sección Categorías	131
	C. Resumen Entrevista Entel	133
C.1.	Reclamos en Entel	133
C.2.	Modelo de retención de clientes	135
C.3.	Entel y Twitter	136